DOCUMENT RESUME

ED 373 634 HE 027 620

AUTHOR	Schwabe, Robert A.; Cherland, Ryan M.
TITLE	Statistical Packages: Some Bugs and Errors. AIR 1994 Annual Forum Paper.
PUB DATE	May 94
NOTE	22p.; Paper presented at the Annual Forum of the Association for Institutional Research (34th, New Orleans, LA, May 29-June 1 1994).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Computer Software; Computer Software Development; Computer Software Evaluation; Debugging (Computers); Higher Education; *Institutional Research; *Statistical Analysis; Statistical Data
IDENTIFIERS	*AIR Forum; Biomedical Computer Programs; Internet; Minitab II Programing Language; *Statistical Package for the Social Sciences; Student Administration System

ABSTRACT

This paper looks at bugs and errors in the following commonly used statistical packages: SAS, Statistical Package for the Social Sciences (SPSS), Biomedical Data Plan (BMDP), and Minitab. A section on using the Internet to keep up to date on these packages covers subscribing to statistical software lists, searching statistical software lists, using gopher sites of archived messages, and receiving information directly from SAS. A brief section on the literature notes that journals and magazines publish reviews of statistical software. A following section offers seven examples of bugs and errors associated with either SAS or SPSS which also have relevance to many statistical packages. (Contains 15 references.) (JB)

Statistical Packages: Some Bugs and Errors

Robert A. Schwabe

Director

Office of Institutional Research

California State University, San Bernardino

5500 University Parkway

San Bernardino, CA 92407-2397

(909) 880-5052

Ryan M. Cherland

Principal Analyst

Office of Institutional Research and Planning

University of Kansas

P.O. Box 505

Lawrence, KS 66044-0505

(913) 864-4412

Paper presented at the

Thirty-Fourth Annual AIR Forum

New Orleans, Louisiana

May 29 - June 1, 1994

Running Head: STATISTICAL PACKAGES: SOME BUGS AND ERRORS

for Management Research, Policy Analysis, and Planning

This paper was presented at the Thirty-Fourth Annual Forum of the Association for Institutional Research held at The New Orleans Marriott, New Orleans, Louisiana, May 29, 1994. This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC Collection of Forum Papers.

Jean Endo
Editor
Forum Publications

Abstract

Most offices of institutional research employ one or more of the common statistical packages: SAS, SPSS, BMDP, or Minitab, for its reporting and analytical work. While textbooks and instructors often note the bugs and errors that occur in a given application or version of the software, the practitioner often overlooks these cautions. This paper: 1) addresses these issues with both historical and technical perspectives; 2) provides examples which are of interest to institutional researchers; and 3) offers techniques for updating one's knowledge of bugs and errors in specific software packages using the resources of electronic listservers and special interest groups.

Statistical Packages: Some Bugs and Errors

Introduction

Since most offices of institutional research employ one or more of the common statistical packages: SAS, SPSS, BMDP, or Minitab, for their reporting and analytical work, we have sought to address the issues associated with finding bugs and errors in these and other computer statistical packages. We will provide some examples which are of interest to institutional researchers and offer techniques for updating ones knowledge of these issues.

As bugs and errors are reported, corrections are being made as either fixes, workarounds or new versions. Also, packages are being ported to new platforms which sometimes introduce new bugs and errors. Further, new and revised software packages, algorithms, statistical approaches, operating systems and machines are continually coming into being. Consequently, it is virtually impossible to keep up with the changes. However, we hope to suggest some useful approaches. Our two basic sources of data for this paper have been: a) Internet archives and current discussion groups; and b) the reviews and comments found in professional journals and other periodicals, as well as conversations with colleagues.

The importance of this topic is emphasized by Marshall (1992) as he reports the on-going debates which have been published as pamphlets, discussed over internet on Stat-L, and

noted in various other publications. In addition to allegations of software plagiarism between SYSTAT's Leland Wilkinson and StatSoft's Pawel Lewicki, the debate has reminded the users of statistical software where he notes that:

> According to William Eddy, a statistics professor at Carnegie-Mellon University in Pittsburgh and chairman of the National Research Council's Committee on Applied and Theoretical Statistics, leaders in the field have been trying for 15 years to establish guidelines for computer programs, and it has been hard to reach a consensus, because every program has its flaws and its devotees. (p. 153)

The problems in the software often do not concern users of these packages who do not always realize what they are asking for or what they are getting. As computers become faster and easier to user, the chances of misuse grow. As Searle puts it (1989, p. 189), "...this great utility can be intelligently realized if, and only if, we really do know (not just think we know, but actually know) what it is that is being computed." One of the ways for the users of these packages to stay knowledgeable is by keeping in contact with the most expert users of these systems. The quickest and easiest method to do this is through the Internet.

Internet as a Resource

The Internet is a rich resource for information on just

about anything. Fraase (1994, p. 10) describes the Internet as:

> ...a collection of high-speed networks composed of the national backbone network provided by the National Science Foundation and a hierarchy of more than 5,000 attached regional, state, federal agency, campus and corporate networks...There are more than 750,000 computers and workstations of various sizes connected to the Internet with millions of users.

Internet resources will allow one to collect information about problems in statistical packages quickly. There are several methods to access sites which provide information about statistical software packages. What would work for you depends on what you have available at your institution and your level of comfort with different types of software. One thing that most individuals have available to them is an e-mail address. With an e-mail address you can subscribe to those discussion lists related to statistical software that you use regularly. Not only will you hear about any programming bugs which users may find, but you will also learn more about programming in that specific software.

Subscribing to Statistical Software Lists

To subscribe to the discussion lists of the common packages you need to send a message to the LISTSERV where the discussion list is located. For most lists send e-mail to listserv@<address> with a one line message:

subscribe <list> <yourfirstname> <yourlastname>

For instance, if Jane Doe wanted to subscribe to SAS-L she would send mail to LISTSERV@UGA.CC.UGA.EDU and put the following line in her e-mail message:

SUBSCRIBE SAS-L Jane Doe

Additional e-mail lists for the common statistical packages and how to subscribe to them are listed below.

Descriptive name	List	Address to send subscription
BMDP Software Users	bmdp-l	listserv@vm1.mcgill.ca
Minitab list	minitab	mailbase@mailbase.ac.uk
SAS list	sas-l	listserv@uga.cc.uga.edu
SPSS list	spssx-l	listserv@uga.cc.uga.edu
Statistical list	stat-l	listserv@vm1.mcgill.ca

A detailed explanation of the statistical discussion groups and other related internet resources can be found in an e-mail article titled Internet Resources for Statistics, by Una Smith (1994, February 27) of Yale University (available from the gopher at jse@stat.ncsu.edu, EdStat-L archives, log9402, article Draft statistics FAQ).

Once you are subscribed to a list you can monitor the articles for useful information and concerns of the software users. You should also note that the traffic on these mail lists can be quite heavy so if you do subscribe be prepared for a high volume of mail. For those who would prefer to use the Usenet services, the SAS and SPSS lists are mirrored on Usenet

discussion groups, comp.soft-sys.sas for SAS discussion, and comp.soft-sys.spss for SPSS discussion.

You can post a message to a list if you have any questions concerning a statistical package and its use. When you send a mail message to the list you would address it not to the LIST-SERV but to the list itself. If Jane Doe wanted to ask about a problem she was having in SPSS she would send her message to SPSSX-L@UGA.CC.UGA.EDU, and it would be broadcast to all the subscribers of the list. A great way to get quick and useful answers from people who are experts in using the various soft-ware packages. As of January 1994 (Smith, 1994) it was estimated that 29,000 people read the comp.soft-sys.sas discus-sion group and 27,000 people read the comp.soft-sys.spss group, so it is likely you would receive a response to your request.

Searching Statistical Software Lists

If you have a question about a particular software package that you feel may have been asked previously, you can do what is called a data base search. LISTSERV managers have built in text search and retrieval capabilities which allow you to search the archives of a discussion group for key words, phrases, specific authors, and even within certain time periods. To find out more information about this capability send an e-mail message to a LISTSERV site (a sure one is LISTSERV@BITNIC) with the following line in its body: INFO DATABASE. You should receive back a file called LISTDB MEMO which will contain detailed information about

the commands available for this type of searching.

An example of an e-mail message used to search a LISTSERV discussion list archive is shown below. In this situation, the query is going to request that the SAS-L discussion group archives be searched for any references to the term 'bug' in the messages.

```
To:  LISTSERV@UGA.CC.UGA.EDU
//BUGS  JOB ECHO=NO
DATABASE SEARCH DD=RULES
//RULES DD *
SEARCH BUG IN SAS-L
INDEX
/*
```

The resulting output is a list of the subject lines of all the articles which contained the term 'bug'. A partial listing of the output is provided below.

```
 > SEARCH BUG IN SAS-L
--> Database SAS-L, 204 hits.

> INDEX
Item #   Date    Time  Recs   Subject
------   ----    ----  ----   -------
000080 93/10/05 07:57   23   Debugging a .SCL Entry
000085 93/10/05 09:55   26   .SCL Entry Problem Solved
000088 93/10/05 19:25   34   EMITS
000089 93/10/05 15:43   51   PROC SQL ALTER TABLE
```

BUG....MVS R6.08

...

004596 94/05/17 17:21 27 Re: SAS Bug or Feature

004625 94/05/18 07:02 34 Re: Tape drives

004629 94/05/18 10:46 34 How the (tape) worm turns!

004659 94/05/19 10:00 25 Tape Access in SAS

Not all of the articles will be relevant to concerns about a programming bug. This is because a string such as bug may be found as a part of many other words. However, from the subject line you should be able to make a guess about its relevance to your interests.

Gopher sites of archived messages

If you have access to telnet or gopher software you can gopher to an archive of discussion list articles for SAS, SPSS, and STAT-L. These archives are rich sources of previous mail traffic and are found at jse.stat.ncsu.edu. If you have gopher software you can type gopher jse.stat.ncsu.edu. If you need to telnet to the site you would enter telnet jse.stat.ncsu.edu and then login as guest when prompted. You will then be presented with a menu of choices. For the purposes of this paper, you would select the 7th option Other Discussion Groups, and then be presented with a submenu. This submenu contains three primary items of interest, the SAS-L Discussion Archives (articles dating back to November 1986), the Stat-L Discussion Archives (articles dating back to December 1989), and the SPSSx-L Dis-

cussion Archives (articles dating back to July 1992). When you select your discussion group of interest you will be able to perform a search on all of the stored articles or a more recent subset of the articles. After all the articles containing the phrase of interest are found you can read the articles and then mail, save, or print those articles that you find of particular interest.

Another gopher site is the StatLib gopher at lib.stat.cmu.edu. It presents gopher access to the StatLib archives. These archives contain statistical software, data sets, directories of statisticians, the collection of Applied Statistics algorithms, as well as many other resources.

Receiving Information directly from SAS

SAS maintains its own technical support list called TSNEWS-L. This is not a discussion list but a list which broadcasts any important updates, fixes, or new problems with the SAS system. If you are a user of SAS software it would be worthwhile to subscribe to the service. To subscribe send the following message to VM.SAS.COM:

SUBSCRIBE TSNEWS-L yourfirstname yourlastname

In addition, you can also send a data base search query as previously discussed to find out any particular problems regarding specific procedures with the software. The notices are quite useful in that they provide information about what operating system platforms are effected and whether or not there are any

fixes or zaps for the problem. This type of information service provides users of SAS with up-to-date information regarding the software and any problems associated with it.

Literature as a Resource

As the current editor of the section entitled Statistical Computing in The American Statistician, Richard Goldstein (1992) lists some twenty-three journals and magazines that regularly publish reviews of statistical software. In fact, he indicates that he will provide a diskette of some 300 reviews taken from these and other publications to interested readers. His list includes such journals as: British Journal of Mathematical and Statistical Psychology, Journal of Applied Econometrics, Journal of Marketing Research, and The American Statistician. The magazines cited include: BYTE, InfoWorld, MacUser, and PC Magazine. Further, he notes that several vendors of statistical software have started or will soon start a technical journals.

In the same article as noted above Goldstein (1992) suggested additional procedures and items that he hopes software packages would some day provide. This appears to be a constant concern for many analysts, since some of them write their own programs or algorithms and make them available via various journals, textbooks or other publications. For example, Applied Statistics, a Journal of the Royal Statistical Society has been publishing statistical algorithms since 1968. Over the years, these various algorithms are often noted, commented upon and

sometimes revised. The book by Griffiths and Hill (1985) is a compilation and a discussion of many of them through 1984. At that time 207 had been published. Those with a familiar titles include: AS45 Histogram plotting; AS51 Log-linear fit for contingency tables; AS75 Basic procedures for large, sparse or weighted linear least squares problems; AS111 The percentage points of the normal distribution; and AS169 An improved algorithm for scatter plots. The latest one is AS291 which was published in 1994 volume 43 number 2. Most of these algorithms including more recent ones are available via Internet using FTP retrieval from StatLib at Carnegie Mellon University as noted above and in Newton (1993).

Many textbooks either support the statistical package, Minitab, which is used primarily in teaching or provide especially designed software in the form of accompanying diskettes. Because of some of the bugs and errors noted below, it is our opinion that such software should not be used for professional analysis. This software is primarily designed for teaching and illustrative purposes and may not have been subject to professional and technical scrutiny.

Some Examples of Bugs and Errors

Several examples of bugs and errors are given below as being associated with either SAS or SPSS specifically but have relevance to many statistical packages.

Example 1

A recent study by Uyar and Erdem (1990) found that it was possible to obtain widely different summary statistics from the regression procedure in SAS. They stated that the procedure did not necessarily cross-check all the options in the model statement against the specified equation or the data. Their paper illustrated three cases of modeling in PROC REG which dealt with forcing the intercept through zero. When this was done using the PROC option of NOINT (no intercept), the results were quite wrong. When the program was rerun using the RESTRICT statement and not the NOINT option the results were correct. The main point the authors wanted to make was that they felt that the SAS manual was not explicit in its warning about how the R^2 and F statistic were being calculated for the model.

Okunade, Chang, and Evans (1993) followed up the Uyar and Erdem article by running the same no-intercept regression model with other statistical software packages. What they found was that other statistical packages had similar problems calculating R^2 and F with the no-intercept model, most notably, BMDP(version PC90) and SPSS/PC+(version 4.1). The conclusions of Okunade, Chang, and Evans (1993, p. 303) were:

> In view of the findings in this study, researchers are simply advised to be sufficiently adept in statistical/econometric estimation theories and applications and pay closer attention to the "notes"

and/or "warning" statements accompanying their regression program outputs and user's manuals.

It should be noted that the book, SAS System for Regression (Freund & Littell, 1991), published by SAS warns the user that, "even when the conditions implied by the NOINT option are reasonable, the results of the analysis have some features that may mislead the unwary practitioner (p. 33)," and they conclude the section with the comment, "Another way of looking at the no intercept model is afforded by the RESTRICT statement (p. 37)".

Example 2

Kaplon, H. (1993) responded to a bug in SPSS suggested by Jensen on the SPSS-L. The issue was the comparison of pairs of numbers which were expected to be equal but did not test to be so. The answer was explained to be the problem of expressing decimal numbers in binary form and converting back again. Equal expressions are not always obtained. For example, in some cases both 5.999999 and 6 denote the same real number. The comparison is based upon the methods used by FORTRAN uses to handle floating point and integer values. A couple of inequalities were introduced to create a valid comparison methodology and thus emphasize the fact that the various statistical packages are intimately dependent upon hardware and compilers.

Example 3

The issue of round-off errors is discussed by Jean-Francois Colonna (1993) in his article entitled The Subjectivity of Com-

puters, he introduces the notion of computer subjectivity or numerical relativity. He discusses an example of numerical integration which is appropriate for some statistical software as it is executed on three different computers. As the number of iterations increased the results from each of the three machines diverged significantly. He also ran the experiment on four other machines with similar results. Sensitivity to initial conditions were noted as well. These problems are related to the fact that computers express real numbers with rational approximations determined by a finite number of bits such as 32 or 64. He asserts that there are "...problems linked to rounding errors that are indigenous to each machine." (p. 15). His concluding paragraph is:

> Today, when both fundamental and applied research relies so heavily on computers, it is essential that users (engineers, scientists and students) should be perfectly aware of the dangers, and that efforts should be made to find real solutions to this very real problem, if such solutions exist. (p. 18)

Example 4

The current version of SPSSX for Macintosh computers contains a bug which causes the system to crash when the online help routine is being used and the attempt is made to open any other application window on the desktop.

Example 5

Nichols (1992) acknowledges that a bug reported by Paul Jackson via several e-mail postings does indeed exist. It occurs when an oblique rotation is asked for in the factor analysis procedure. The reported regression scores are incorrect since the structure matrix is used instead of the rotated matrix. A workaround is to use the Anderson-Rubin or Bartlett scores since they are the same as the correct regression scores.

Example 6

We note with interest that known errors in SPSS for IBM VM/CMS, Release 4.1 can be obtained by the INFO command, but the items presented may be reproduced by licensees of the SPSS system for local distribution only. Other use without the prior permission of SPSS Inc. is not authorized. The discussion includes both fixed and unfixed errors. A request is included asking users to report unreported unfixed errors to SPSS Technical Support.

Example 7

In his discussion of balanced data, Schwartz (1993) compares the ANOVA output obtained for a balanced mixed model by three mainframe packages: BMDP, 1990 Release; SAS Version 6, 1989 and SPSS Version 4.0, 1990. His careful summary of his analysis reveal significant differences in output, cautions on usage and notes incorrect results associated with each of four runs: BMPD-3V; BMPD-8V; SAS-GLM & VARCOMP; and SPSS-MANOVA. He

notes that "It is unfortunate that some computer packages still report incorrect results, have stringent rules for data input, have awkward syntaxes for the command languages, or are insufficiently flexible to provide a full analysis." (p. 58). He concludes with seven suggestions as to what modern statistical packages should provide.

Conclusion

As we've noticed by the several examples that we have discussed, the problem of finding and documenting bugs and errors in the various statistical packages is very difficult since the scene is constantly changing. New packages are being offered, new versions are appearing and old programs are being ported to other platforms with different hardware, operating systems and compilers.

We have illustrated that all computer software designed for performing statistical analysis is subject to bugs and errors throughout its life cycle even as it becomes revised, updated and ported to other computer platforms. We have noted that there are numerous sources of data for discovering these bugs and errors as well as finding some temporary work-arounds or other solutions. These sources include Internet and some of the many publications that are available to the user as well as the professional statistician.

Finally, we suggest that the user of statistical software packages should continue to exercise caution and care. The often

stated dictum of knowing one's data via some exploratory techniques, fitting them to appropriate statistical models and finally processing them with a well chosen statistical software package should always be followed.

References

Colonna, J. (1993). The Subjectivity of Computers. Communications of the ACM, 36 (8), 15-18.

Fleming, N. S. (1991). An Additional Note on Regression Procedures in SAS. The American Statistician, 45(3), 261.

Fraase, M. (1994). The PC Internet Tour Guide: Cruising the Internet the Easy Way. Chapel Hill, NC: Ventana Press, Inc.

Goldstein, R. (1992). Editor's Notes. The American Statistician, 46(1), 48-49.

Griffiths, P., and Hill, I. D. (Eds.). (1985). Applied Statistics Algorithms. Royal Statistical Society/Ellis Horwood Ltd. London.

Kaplon, H. (1993, January 20). No Bug in SPSS. In SPSSX-L archives. Available from gopher @jse.stat.ncsu.edu.

Freund, R. J., & Littell, R. C. (1991). SAS System for Regression, Cary, NC: SAS Institute Inc.

Marshall, E. (1992). Statisticians at Odds Over Software Ownership. Science, 255, 152-153.

Newton, H. J. (1993). New Developments in Statistical Computing. The American Statistician, 47(2), 146-147.

Nichols, D. (1992, August 7). Re: SPSSX. Factor Analysis. In SPSSX-L archives. Available from gopher @jse.stat.ncsu.edu.

Okunade, A. A., Chang, C. F., & Evans, R. D. (1993). Comparative Analysis of Regression Output Summary Statistics in Common Statistical Packages. The American Statistician, 47(4), 298-303.

Schwartz, C. J. (1993). The Mixed-Model ANOVA: The Truth, the Computer Packages, The Books Part I: Balanced Data. The American Statistician, 47(1), 48-59.

Searle, S. R. (1989). Statistical computing packages: Some words of caution. The American Statistician, 43(4), 189-190.

Smith, U. (1994, February 27). Draft statistics FAQ. In EdStat archives. Available from gopher @jse.stat.ncsu.edu.

Uyar, B., & Erdem, O. (1990). Regression Proc dures in SAS: Problems? The American Statistician, 44(4), 296-301.